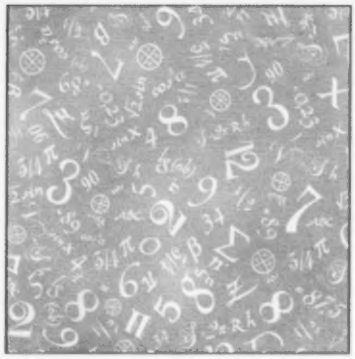
Component of Statistics Canada Catalogue no. 82-003-X
Health Reports

Article

Strategies for handling normality assumptions in multi-level modeling: A case study estimating trajectories of Health Utilities Index Mark 3 scores

by Julie Bernier, Yan Feng and Keiko Asakawa



December, 2011

Statistics Canada Statistique Canada

Canada

Strategies for handling normality assumptions in multi-level modeling: A case study estimating trajectories of Health Utilities Index Mark 3 scores

by Julie Bernier, Yan Feng and Keiko Asakawa

Abstract

Background
With longitudinal data, lifetime health status dynamics can be estimated by modeling trajectories. Health status trajectories measured by the Health Utilities Index Mark 3 (HUI3) modeled as a function of age alone and also of age and socio-economic covariates revealed non-normal residuals and variance estimation problems. The possibility of transforming the HUI3 distribution to obtain residuals that approximate a normal distribution was investigated.

Data and methods
The analysis is based on longitudinal data from the first six cycles of the National Population Health Survey (NPHS). The data pertain to 7,784 individuals, who, in 1994/1995, were aged 40 to 99, were living in private households, and had complete information on HUI3. A multi-level growth model was used to examine the hierarchical structure of NPHS data (repeated measurements nesting within respondents). The transformation of arcsine [2 × (HUI + 0.36) / (1 + 0.36) – 1] was used to improve the distribution of the residuals at both levels and limit the conditional mean to the -0.36 to 1.00 interval. A model was estimated using socio-economic determinants. Analyses were performed with SAS and MLwiN.

Results
After the transformation of HUI3, the model was satisfactory and allowed for inclusion of new socio-demographic and health variables in order to estimate their impact on the health-related quality of life of aging populations. Because of the complex transformation of the arcsine model, the regression coefficients were not interpreted. Instead, the estimation results were summarized graphically.

Keywords
Health status indicators, health surveys, health transition, logistic models, longitudinal studies, multilevel growth model

Author
Julie Bernier (1-613-951-4556; julie.bernier@statcan.gc.ca) and Keiko Asakawa are with the Health Analysis Division and Yan Feng is with the Income Statistics Division at Statistics Canada, Ottawa, Ontario, K1A 0T6.

Longitudinal data from Statistics Canada's National Population Health Survey (NPHS) can be used to assess health status dynamics. For more than a decade, the NPHS collected repeated samples every two years. Estimations of repeated measures data are facilitated by using a growth-curve (multi-level) model approach,[1] which allows the estimation of within-individual (level-1) and between-individual (level-2) variations in outcomes. With a growth-curve model, the dynamics can be presented by a trajectory, and associations between socio-economic and health determinants and trajectories of health-related quality of life (HRQL) can be examined.

As with any regression method, the utility of the estimation results depends on the degree to which model assumptions are met. In single-level models, assumptions about model (for example, linearity, omitted variables, interactions) and stochastic specifications (for example, heteroskedasticity, normality of errors) should be assessed carefully.[2,3] This article focuses on the normality of error assumption in a growth-curve model setting. In such a model, where respondents are considered as the level-2 unit and occasions (time) within each respondent are considered as the level-1 unit, the normality of error assumption indicates univariate normality of residuals at level-1 and univariate or multivariate normality (if more than one parameter was considered as random) of random components at level-2. Failure of the normality assumption at level-1 will not bias estimation of the fixed effects, but it will introduce bias into standard errors at both levels, thereby affecting the validity of confidence intervals and hypothesis tests. Estimation of the level-2 fixed effects will not be biased by non-normality of the errors at level-2.

However, the presence of skewness will affect inferences at level-2.[2]

In analyses of longitudinal data from surveys such as the NPHS, the normality of error assumption must be considered because population health outcomes such as HRQL are often skewed. A measure of HRQL collected in the NPHS is the Health Utilities Index Mark 3 (HUI3). This is a generic, multi-attribute, continuous preference-based indicator that describes health status with a single summary measure ranging from -0.36 to 1.00 (1.00 = perfect health; 0.00 = dead; -0.36 = a state worse than dead).[4] However, because of the highly skewed distribution of HUI3,[5,6] the normality of error assumption may be violated when it is used in estimating growth curve models. When determinants of health variables are introduced into the model, standard errors of the parameters cannot be estimated properly.

One method of dealing with violation of the normality assumption is to transform the outcome variable to improve the error distribution. (The transformation is intended to yield unskewed residuals, not unskewed dependent variables.) This study assesses the utility of *arcsine* transformation, which stabilizes the variance and improves the symmetry of the residuals.[7] An earlier study showed that when an untransformed HUI3 was used as the outcome variable, the predicted HUI3 scores fell below the theoretical lower bound of -0.36.[1]

A preliminary investigation of the arcsine transformation of a particular form (arcsine[2 × (HUI3 + 0.36) / (1 + 0.36) -1]) resulted in predicted back-transformed HUI3 scores that were above the theoretical lower bound of -0.36. While the arcsine transformation has been used,[8] to our knowledge it has not been applied to analyses of longitudinal population health data. Thus, how this transformation handles the normality of error assumption in a growth-curve model setting is not known.

The primary objective of this study was to evaluate the feasibility of using the arcsine transformation from the family of trigonometric functions to estimate growth-curve models with non-normally distributed residuals. It was assessed for a simple socio-economic model that included marital status, education and household income. Two other transformations were also considered: one from a log family (natural logarithmic transformation) and another from an exponent family (square-root transformation). The performance of these three transformations was compared.

An additional challenge of the model transformation is interpretation of estimation results. Because back-transformation of estimated coefficients is difficult, if not impossible, the secondary objective was to present a graphical approach to interpreting estimation results, based on a model with a transformed dependent variable.

A case study focusing on the performance of growth-curve models using various types of transformations of HUI3 as an outcome variable was conducted. The aim was to provide a pragmatic approach to handling the normality of error assumption, not to find the best-fitting model among possible types of functional forms, estimation techniques, or model specifications. This study demonstrates the potential of the arcsine transformation in a growth-curve model setting by comparing its performance with those of other commonly used transformation methods.

Methods

Data source

The data are from the household component of the 1994/1995 to 2006/2007 National Population Health Survey (NPHS), which collected longitudinal information about Canadians' health and socio-demographic characteristics. The target population was household residents in the ten provinces in 1994/1995, excluding residents of Indian Reserves and Crown Lands, health institutions and some remote areas in Ontario and Quebec; full-time members of the Canadian Forces; and all residents (military and civilian) of Canadian Forces bases.

To study HUI3 trajectories, data from the NPHS longitudinal square file were used. The square file includes all 17,276 respondents to cycle 1 (1994/1995), regardless of their response pattern in the next six cycles. The longitudinal sample size remained the same for all cycles.[9,10]

For this study, respondents aged 40 to 99 in 1994/1995 who had complete HUI3 information were selected. Exclusion of the small number of people aged 100 or older did not affect the estimates of parameters in the regression model. The target population included the 252 respondents who were institutionalized at some point during the six follow-up cycles. The 1,295 respondents who died during follow-up were also included, but only for the cycle in which their death was reported; information for subsequent cycles was left as missing. The final sample consisted of 7,784 respondents.

Outcome variable

The outcome variable was HUI3, a continuous variable that ranges from -0.36 to 1.00 (1.00 = perfect health; 0.00 = dead; -0.36 = a state worse than dead). To compare models with various transformations, the HUI3 scores were transformed as described below.

Independent variables

Linear and non-linear forms of a variable indicating the age of respondents were included in the model to represent time in the analyses. The age variable was centered at 57, the mean age at baseline. Gender, marital status, education and household income were added as independent variables. Gender (female as reference group) was included as a time-invariant variable. Marital status, education and household income were included as time-varying covariates. Marital status was categorized as married/common-law/living with partner (reference group) or single/separated/divorced/widowed. Education was categorized as less than secondary graduation or at least secondary

graduation (reference group). Household income was categorized as low (less than $15,000), middle ($15,000 to $29,999), or high ($30,000 or more; reference group).

Two sets of time-varying dummy variables were created as control variables: place of residence (1 if institutionalized, 0 otherwise) and the state of being dead (1 if dead, 0 otherwise). To account for mortality effects in the analyses,[1] the record of the first report of death was retained in the analyses by assigning a value of HUI3 = 0.00 to the dependent variable. For the independent variables, the last observed value for each was assigned to the first record of death; subsequent records were left as missing.

Modelling

The NPHS data consist of repeated measurements of respondents over six cycles of data collection. The hierarchical structure of the data—repeated measurements nesting within respondents—can be modeled using a two-level growth model. A multi-level growth model simultaneously incorporates within-person and between-person change. The within-subject model (level-1) was specified as a function of a set of growth parameters and individual time-varying characteristics over time with measurement error. The growth parameters and time-invariant individual characteristics, which are a source of heterogeneity, were specified in the between-subject model (level-2) to capture the variation of growth across the population.

As in an earlier study,[1] age was expressed in a cubic form: a linear growth rate (Age), a quadratic growth rate (Age^2), and a cubic growth rate (Age^3). Only the intercept was considered as varying randomly across individuals, which is a simpler form than that presented in the previous study.[1]

The growth model including household income, education, marital status, death and institutionalization at level-1, and gender at level-2, is:

Level-1:

$$Y_{ij} = \beta'_{0j} + \beta'_{1j}\, Age_{ij} + \beta'_{2j}\, Age_{ij}^2 + \beta'_{3j}\, Age_{ij}^3 + \sum_{n=4}^{N} \beta'_{nj} X_{nj} + \varepsilon'_{ij}$$

$$\varepsilon'_{ij} \sim N(0, \sigma'^2_{\varepsilon})$$

Level-2:

$$\beta'_{0j} = \beta'_{00} + \beta'_{10}\, \text{Gender}_j + \zeta'_{0j}$$

$$\zeta'_{0j} \sim N(0, \sigma'^2_0)$$

where Y_{ij} is the HUI3 scores for individual j at cycle i. The level-1 4 β's are the true growth parameters varying across individuals. σ_{ε}^2 describes within-individual random deviation of HUI3 scores from his/her own trajectory. β_{00} represents the population mean HUI3 score at age 57, and σ_0^2 represents the between-individual random deviation in mean HUI3 scores (at age 57).[3]

A person-period dataset was created for the analysis of multi-level growth models. The date of birth and the date of the interview recorded in the NPHS microdata made it possible to use the time-unstructured characteristic of the NPHS data (the actual ages of respondents may not necessarily change by a two-year increment between assessment periods) by calculating respondents' actual age.[3] Therefore, the model was fit using the actual numeric values of AGE (the difference between the interview date and the self-reported date of birth) as a temporal variable, rendering the person-period data time-unstructured. As in most longitudinal studies, the data were unbalanced because of attrition. Growth curve models allow the estimation using the time-unstructured and unbalanced data. Of the 7,784 individuals in the sample, 2,989 (38.4%) had six records; 1,546 (19.9%) had five records; 1,141 (14.7%) had four records; 928 (11.9%) had three records; 748 (9.6%) had two records; and 432 (5.6%) had one record.

Transformation

Arcsine transformation was used to transform the dependent variable (HUI3) to improve the normality of residuals. The arcsine transformation is usually applied to a variable with a [-1, 1] range. If a variable X ranges from -1 to 1, arcsine(X) will range from negative infinity to positive infinity. However, HUI3 is bounded by -0.36 and 1.00. Therefore, the arcsine transformation may not be implemented effectively to modify the HUI3 distribution because of the theoretical lower bound of -0.36. To facilitate an arcsine transformation, HUI3 was first linearly transformed so that the transformed HUI3 scores were bounded by -1 and +1 using the following equation:

$$2 \times \frac{HUI3 + 0.36}{1 + 0.36} - 1$$

The arcsine transformation was implemented using the transformed HUI3 scores (arcsine [2 × (HUI3 + 0.36) / (1 + 0.36) – 1]). This improved the prediction of the trajectory for an aging population by allowing the predicted back-transformed HUI3 scores to lie within the theoretical lower bound of -0.36. By contrast, in the earlier study, the predicted HUI3 scores fell beyond -0.36 when an untransformed HUI3 was used as the outcome variable.[1]

Assessment of normality assumption

The normality of error assumption was assessed by comparing skewness statistics among alternative models. Higher-level statistics like kurtosis were not estimated. In these analyses, errors were considered close to normally distributed if the skewness statistic was zero.[11] An improvement in the normality of error assumption is considered to have occurred if the skewness statistics of level-1 and level-2 errors for one model are closer to zero than those of another model. Normal probability plots were also used to assess the distribution of residuals at both levels. Straight-line plots of theoretically generated normal scores against standardized residuals indicate normally distributed residuals.[12]

The appropriateness of the arcsine transformation was assessed by comparing distributions of residuals of the model with those of the untransformed model and two alternative

models, each based on a different form of transformation of the dependent variables: natural logarithm[13] and square-root.[7] The normality of error assumption was assessed by comparing skewness statistics and normal probability plots across the four models.

Interpretation of models

Based on results of the arcsine-transformed model, a graph representing estimated trajectories was constructed by setting values of explanatory variables at different levels. Specifically, trajectories of back-transformed HUI3 scores were plotted by gender, marital status, education, household income, and place of residence (community or institution).

All analyses were performed with SAS and MLwiN. Models were weighted using the sampling weights to account for the unequal selection probabilities of the NPHS, and the weights were applied to the second level of the model. Variance estimates were not adjusted for the complex sampling design of the survey.

Results

Preliminary descriptive investigation of health trajectories

To identify a suitable functional form for the level-1 sub-model and to summarize how health status changes over time, empirical growth plots for HUI3 with smooth nonparametric trajectories were examined by age group. The trajectory of mean HUI3 by age suggested that the level-1 sub-model was nonlinear.[1] The trajectory of HUI3 displays a quadratic and a cubic trajectory (plots not shown). Instead of selecting a unique polynomial form for each person, the highest order polynomial was selected to summarize individual change for any person. Therefore, linear, quadratic and cubic terms of Age were included in all models.

Regression results

Because of the complexity of back-transforming estimated coefficients in the arcsine HUI3 model to the original scale, the estimated parameters are not interpreted separately. Moreover, when a transformation is applied to the outcome variable, the estimated parameters of the transformed model have the least squares properties with respect to the transformed observations only, not with respect to the original observations.[14] Therefore, only parameters of the transformed models are presented in this study.

Comparison of proposed models

Comparisons among models showed that the model with arcsine transformation (Model 3) had the skewness of level-1 residuals as -0.74, which was the closest to zero among the models (Table 1). In particular, the skewness statistics of level-1 residuals in log (Model 1) and square-root models (Model 2) were as large as -1.77 and -1.88, respectively. Comparisons of normal probability plots for untransformed HUI3 (Figure 1a) with arcsine transformation showed that the residual plots for the arcsine model (Model 3, Figure 1b) appeared to be the closest to linearity among the four models (figures for Models 1 and 2 not shown).

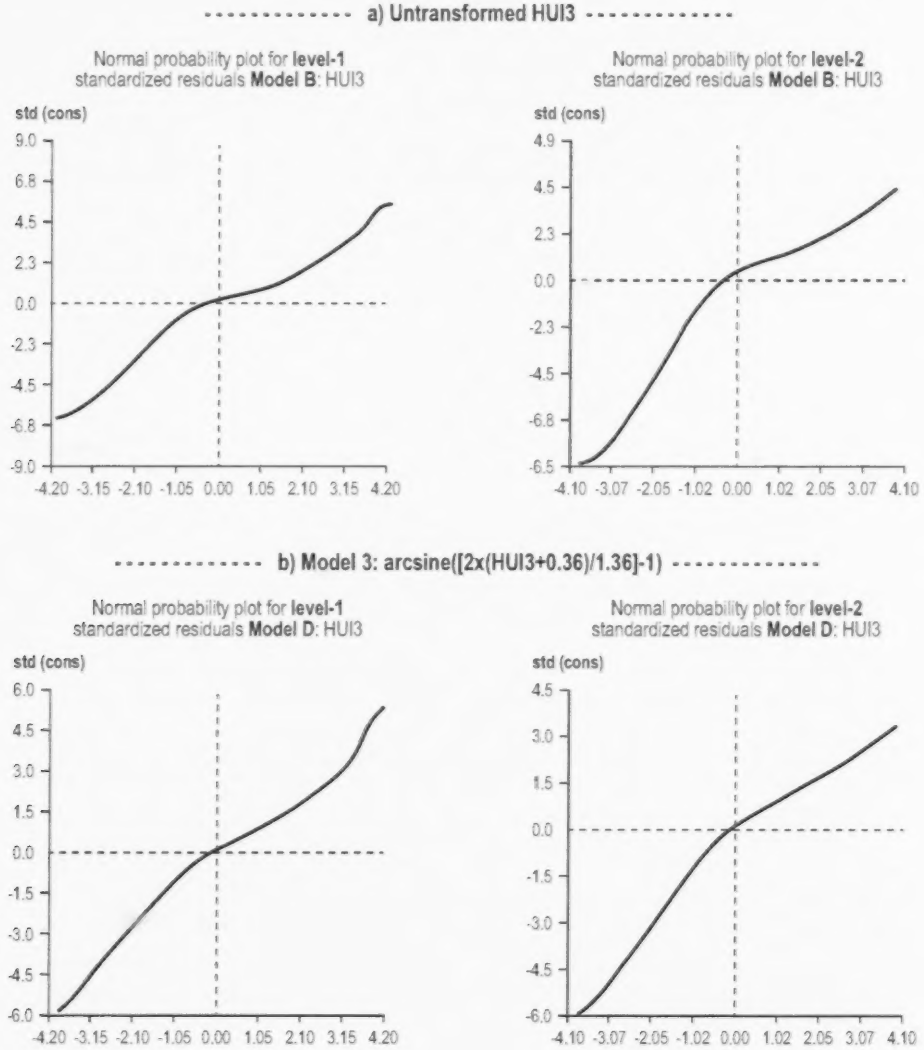
The skewness statistics of level-2 residuals in both Model 1 and Model 2 were approximately -1.90, further from zero compared with level-1. The skewness statistic for the arcsine transformation (Model 3) was -1.06, which was closest to zero among the models. The normal probability plots for standardized residuals at level-2 showed that the plots for Model 3 were the closest to linearity among the four models, although the tails still deviated from normality at the upper end of the distribution (Figure 1b).

Table 1
Comparison of fitting alternative polynomial change trajectories

	Untransformed HUI3	Model 1	Model 2	Model 3
Fixed effects				
Intercept	0.882576**	0.627531**	0.952846**	1.076387**
Mean age	-0.001724**	-0.000962**	-0.000682**	-0.005272**
Mean age squared	-0.000050**	-0.000033**	-0.000023**	-0.000018**
Mean age cubed	-0.000004**	-0.000003**	-0.000002**	-0.000008**
Deceased	-0.668605**	-0.489359**	-0.340475**	-1.134217**
Institutionalized	-0.459094**	-0.352409**	-0.253395**	-0.810786**
Female†				
Male	0.0006329	0.003433	0.002511	0.023969*
Secondary graduation or more				
Less than secondary graduation†	-0.037715**	-0.023636**	-0.016734**	-0.082577**
Married/Common-law/Living with partner†				
Single	-0.01384	-0.007427	-0.005187	-0.03341
Divorced/Separated/Widowed	-0.014254**	-0.00808**	-0.005672**	-0.030819**
High income				
Middle income	-0.020276**	-0.013019**	-0.009237**	-0.042748**
Low income	-0.045075**	-0.029279**	-0.020741**	-0.091397**
Random effects				
Level-1 (σ_e^2 (within-person))	0.024419**	0.010790**	0.00542**	0.106314**
Level-2 (σ_{z0}^2 (intercept))	0.016104**	0.006713**	0.003357**	0.069918**
Skewness				
Level-1	-1.35	-1.77	-1.88	-0.74
Level-2	-1.59	-1.91	-1.93	-1.06

* significantly different from estimate for reference group ($p<0.0005$)
** significantly different from estimate for reference group ($p<0.0001$)
† reference category
.. not applicable
Notes: Untransformed: HUI3 as a dependent variable
Model 1: ln (HUI3+1) as a dependent variable
Model 2: Square-root of [(HUI3+0.36)/1.36] as a dependent variable
Model 3: Arcsine of [2 × (HUI3 + 0.36) / (1 + 0.36) – 1] as a dependent variable
Source: 1994/1995 to 2006/2007 National Population Health Survey, longitudinal square file

Figure 1
Comparison of normal probability plots of residuals

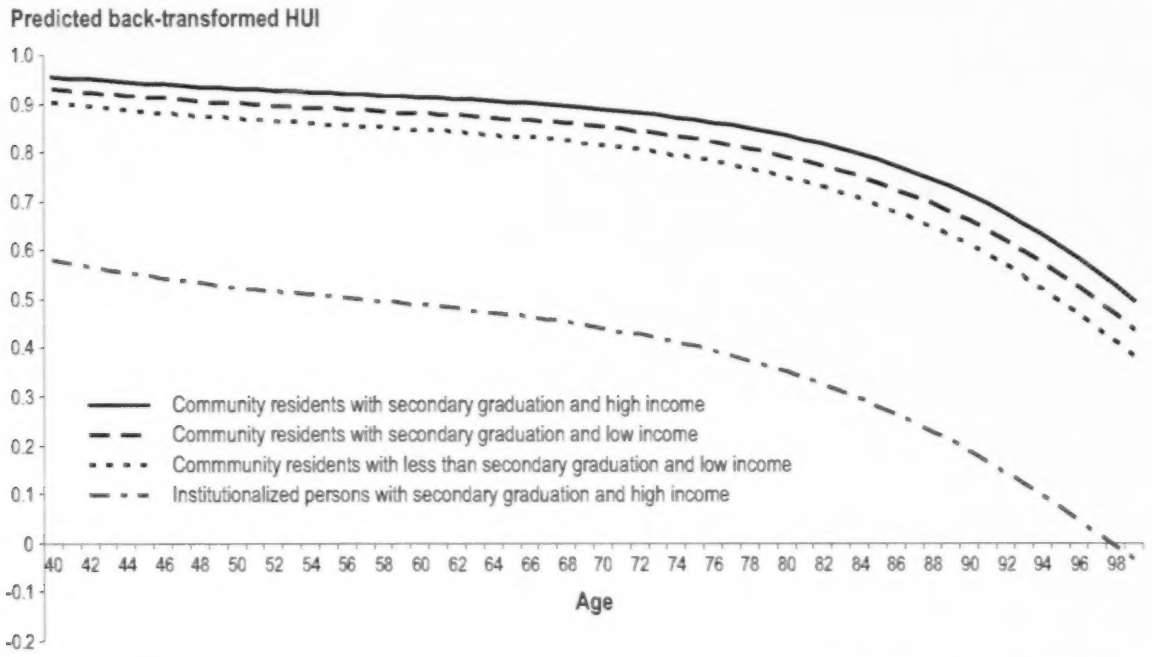


Source: 1994/1995 to 2006/2007 National Population Health Survey, longitudinal square file.

Model interpretation: A graphical approach

As an illustration, the health trajectories for women with selected socio-demographic profiles based on the model with arcsine transformation (Model 3) are presented (Figure 2). When all other socio-demographic characteristics were held constant, women living in the community had substantially different health trajectories than did women in institutions. Women in the community were, on average, much healthier, a difference that became more pronounced with advancing age. Education was also associated with health trajectories. Among women living in the community in a low-income household, those with less than secondary graduation had a lower HUI3 trajectory than did those who were at least secondary graduates. Household income was also related to the variation in health trajectories.

Figure 2
Predicted HUI3 trajectories for four groups of married women aged 40 or older, by single-year-of-age, education and household income, Canada excluding territories, 1994/1995 to 2006/2007


Predicted back-transformed HUI
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Community residents with secondary graduation and high income
Community residents with secondary graduation and low income
Commmunity residents with less than secondary graduation and low income
Institutionalized persons with secondary graduation and high income
40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98
Age

Source: 1994/1995 to 2006/2007 National Population Health Survey, longitudinal square file.

Discussion

In assessing the health status of a population, it is not uncommon for continuous measures of HRQL to be non-normally distributed, leading to the violation of normality assumptions in regression analyses. The issue affects the estimation of growth curve models, in that failure of the normality assumption will bias standard errors, thereby affecting the validity of confidence intervals and hypothesis tests. This analysis presents a case study to determine if transformation of the outcome variable as *arcsine* [2 × (HUI3 + 0.36) / (1 + 0.36) – 1] has the potential to address this problem. Results showed that the arcsine-transformed model reduced the skewness of the residual distribution. The symmetry of the residual distribution was noticeably improved, compared with the untransformed models and models with natural logarithm and square-root transformations. A graphical approach is also presented by plotting predicted back-transformed HUI3 trajectories, although the complexity of interpreting estimation results based on a model with a non-linearly transformed dependent variable is recognized.

The case study is unique in that, to our knowledge, arcsine transformation has not been applied to a growth-curve model in the analysis of population health surveys, and the performance of the transformation was superior to untransformed or natural logarithmic and square-root transformations. The graphical approach is a straightforward way of interpreting results from the arcsine model without the complexity of exploring the back-transformation of the set of estimated coefficients.

Several considerations should be noted. First, the record of death of a respondent was assigned the last observed values for the explanatory variables. This simple approach may not be optimal, as other methods are available.[15] Nonetheless, fewer than 1% of all the records used in the analyses that corresponded to the first record of death were imputed; values for the subsequent cycles were left as missing, so any bias resulting from this approach is likely to be minimal. Second, the three models that were tested are not an exhaustive representation of potential transformations. Third, some measures of HRQL, including HUI3, are subject to a ceiling effect. However, the ceiling effect is unlikely to be an issue in this study because only 10% of all records for respondents aged 40 or older (fewer than 1% of those for respondents aged 65 or older) had a HUI3 score of 1.00 (perfect health). Fourth, the explanatory variables in the models were chosen only to illustrate the usefulness of transforming HUI3. Before choosing a definitive model, more attention must be paid to selecting covariates and examining a possible random slope effect. Fifth, in theory, numeric utilities such as HUI3 are unique

up to positive linear transformation,[16] but the arcsine transformation is non-linear. Nonetheless, the proposed method is pragmatic and useful, with the back-transformed HUI3 trajectories helping to visualize important heterogeneities in health trajectories.

The case study showed the arcsine transformation to be a statistically appropriate way of handling the normality of error assumption in multi-level modeling. It is also useful in estimating the impact of various determinants of health on health trajectories. The method is accessible through most statistical packages and can describe variations in health trajectories among socio-economic groups. The approach can also be used to assess other determinants of health. This case study is an initial attempt to introduce this transformation in a multi-level model setting. Further investigation is warranted to examine the potential of the arcsine transformation for other types of transformations, estimation methods and populations. ■

References

1. Orpana H, Ross N, Feeny D, et al. The natural history of health-related quality of life: a 10-year cohort study. *Health Reports* 2009; 20(1): 29-35.
2. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Second Editon. Thousand Oaks, California: Sage Publications, 2002: 266-7.
3. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence*. Toronto: Oxford University Press, 2003.
4. Feeny D, Furlong W, Torrance GW, et al. Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 System. *Medical Care* 2002; 40(2): 113-28.
5. Eng K, Feeny D. Comparing the health of low income and less well educated groups in the United States and Canada. *Population Health Metrics* 2007; 5(10): 103.
6. van Doorslaer E, Jones AM. Inequalities in self-reported health: Validation of a new approach to measurement. *Journal of Health Economics* 2003; 22: 61-87.
7. Kleinbaum D, Kupper L, Muller K, Nizam A. *Applied Regression Analysis and Other Multivariable Methods*. Third edition. Toronto: Duxbury Press, 1997: 252.
8. Hertley A, Ho KF, McConkey C, Geh JI. Pathological complete response following pre-operative chemoradiotherapy in rectal cancer: analysis of phase II/III trials. *The British Journal of Radiology* 2005; 78: 934-8.
9. Swain L, Catlin G, Beaudet M. Sample design of the National Population Health Survey: its longitudinal nature. *Health Reports* 1999; 10(4): 69-82.
10. Statistics Canada. *National Population Health Survey Household Component: Cycles 1 to 7 (1994/1995 to 2006/2007) Longitudinal Documentation*. Ottawa: Statistics Canada, 2007.
11. Bickel PJ, Doksun KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Oakland, California: Holden-Day, 1977: 388.
12. Johnson RA, Bhattacharyya GK. *Statistics, Principles and Methods*. Toronto: John Wiley & Sons Inc., 1992.
13. Manning WG, Mullahy J. Estimating log models: Transform or not to transform? *Journal of Health Economics* 2001; 20 (4): 461-94.
14. Neter J, Wasserman W, Kutner MH. *Applied Linear Statistical Models*. Second Edition. Homewood, Illinois: Richard P. Irwin, Inc., 1985: 141.
15. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Second Edition. Hoboken, New Jersey: Wiley, 2002.
16. Jehle GA, Reny PJ. *Advanced Microeconomic Theory*. Reading, Massachusetts: Addison-Wesley, 1998.